During post-training, we developed novel training techniques incorporating reasoning capabilities and curated a diverse set of engineering tasks, with the aim to equip Gemini with effective problem-solving skills crucial for addressing modern engineering challenges. Key applications demonstrating these advancements include IDE functionalities, code agent use cases for complex, multi-step operations within full repositories, and multimodal, interactive scenarios such as end-to-end web and mobile application development. Collectively, these efforts have yielded broad and significant improvements in Gemini's coding capabilities. This progress is evidenced by superior performance on established benchmarks: performance on LiveCodeBench (Jain et al., 2024) increased from 30.5% for Gemini 1.5 Pro to 74.2% for Gemini 2.5 Pro, while that for Aider Polyglot (Gauthier, 2025) went from 16.9% to 82.2%. Performance on SWEBench-verified (Chowdhury et al., 2024; Jimenez et al., 2024) went from 34.2% to 67.2%, see Table 3 and Figure 5 in Section 3.2. Furthermore, Gemini 2.5 Pro obtained an increase of over 500 Elo over Gemini 1.5 Pro on the LMArena WebDev Arena (Chiang et al., 2024; LMArena Team, 2025), resulting in meaningful enhancements in practical applications, including UI and web application development (Doshi, 2025a), and the creation of sophisticated agentic workflows (Kilpatrick, 2025).

Factuality

Within the context of generative models, ensuring the factuality of model responses to information-seeking prompts remains a core pillar of Gemini model development. With Gemini 1.5, our research was concentrated on enhancing the model's world knowledge and its ability to provide answers faithfully grounded in the context provided within the prompt. This effort culminated in the December 2024 release of FACTS Grounding (Jacovi et al., 2025), now an industry-standard benchmark for evaluating an LLM's capacity to generate responses grounded in user-provided documents. With Gemini 2.0 and 2.5, we have significantly expanded our scope to address multimodal inputs, long-context reasoning, and model-retrieved information. At the same time, the landscape and user expectations for factuality have evolved dramatically, shaped in part by Google's deployment of AI Overviews and AI Mode (Stein, 2025). To meet these demands, Gemini 2.0 marked a significant leap as our first model family trained to natively call tools like Google Search, enabling it to formulate precise queries and synthesize fresh information with sources. Building on this, Gemini 2.5 integrates advanced reasoning, allowing it to interleave these search capabilities with internal thought processes to answer complex, multi-hop queries and execute long-horizon tasks. The model has learned to use search and other tools, reason about the outputs, and issue additional, detailed follow-up queries to expand the information available to it and to verify the factual accuracy of the response. Our latest models now power the experiences of over 1.5B monthly active users in Google's AI Overviews and 400M users in the Gemini App. These models exhibit state-of-the-art performance across a suite of factuality benchmarks, including SimpleQA for parametric knowledge (Wei et al., 2024), FACTS Grounding for faithfulness to provided documents (Jacovi et al., 2024, 2025), and the Vectara Hallucination Leaderboard (Hughes et al., 2023), cementing Gemini as the model of choice for information-seeking demands.

Long context

Modeling and data advances helped us improve the quality of our models' responses to queries utilizing our one million-length context window, and we reworked our internal evaluations to be more challenging to help steer our modeling research. When hill-climbing, we targeted challenging retrieval tasks (like LOFT of Lee et al., 2024), long-context reasoning tasks (like MRCR-V2 of Vodrahalli et al., 2024), and multimodal tasks (like VideoMME of Fu et al., 2025). According to the results in Table 6, the new 2.5 models improve greatly over previous Gemini 1.5 models and achieve state-of-the-art quality on all of those. An example showcasing these improved capabilities for video recall can be seen in Appendix 8.5, where Gemini 2.5 Pro is able to consistently recall a 1 second visual event out of a full 46-minute video.[2]

Multilinguality

Gemini's multilingual capabilities have also undergone a profound evolution since 1.5, which already encompassed over 400 languages via pretraining. This transformation stems from a holistic strategy, meticulously refining pre- and post-training data quality, advancing tokenization techniques, innovating core modeling, and executing targeted capability hillclimbing. The impact is particularly striking in Indic and Chinese, Japanese and Korean languages, where dedicated optimizations in data quality and evaluation have unlocked dramatic gains in both quality and decoding speed. Consequently, users benefit from significantly enhanced language adherence, responses designed to faithfully respect the requested output language, and a robust improvement in generative quality and factuality across languages, solidifying Gemini's reliability across diverse linguistic contexts.

Audio

While Gemini 1.5 was focused on native audio understanding tasks such as transcription, translation, summarization and question-answering, in addition to understanding, Gemini 2.5 was trained to perform audio generation tasks such as text-to-speech or native audio-visual to audio out dialog. To enable low-latency streaming dialog, we incorporated causal audio representations that also allow streaming audio into and out of Gemini 2.5. These capabilities derive from an increased amount of pre-training data spanning over 200 languages, and development of improved post-training recipes. Finally, through our improved post-training recipes, we have integrated advanced capabilities such as thinking, affective dialog, contextual awareness and tool use into Gemini's native audio models.

Video

We have significantly expanded both our pretraining and post-training video understanding data, improving the audio-visual and temporal understanding capabilities of the model. We have also trained our models so that they perform competitively with 66 instead of 258 visual tokens per frame, enabling using about 3 hours of video instead of 1h within a 1M tokens context window[3]. Two new applications that were not previously possible, but that have been unlocked as a result of these changes are: creating an interactive app from a video (such as a quiz to test students' understanding of the video content) and creating a p5.js animation to show the key concepts from the video. Our recent blog post (Baddepudi et al., 2025) shows examples of these applications.

Gemini as an Agent: Deep Research

Gemini Deep Research (Gemini Team, Google, 2024) is an agent built on top of the Gemini 2.5 Pro model designed to strategically browse the web and provide informed answers to even the most niche user queries. The agent is optimized to perform task prioritization, and is also able to identify when it reaches a dead-end when browsing. We have massively improved the capabilities of Gemini Deep Research since its initial launch in December 2024. As evidence of that, performance of Gemini Deep Research on the Humanity's Last Exam benchmark (Phan et al., 2025) has gone from 7.95% in December 2024 to the **SoTA score of 26.9% and 32.4% with higher compute** (June 2025).

[2] For further discussion on long context capabilities, challenges, and future outlook, the Release Notes podcast episode "Deep Dive into Long Context" provides additional insights and discussion: `https://youtu.be/NHMJ9mqKeMQ`.

[3] This is referred to as low media resolution in the API: `https://ai.google.dev/api/generate-content#MediaResolution`.

2.7. The path to Gemini 2.5

On the way to Gemini 2.5 Pro, we experimented with our training recipe, and tested a small number of these experimental models with users. We have already discussed Gemini 2.0 Flash Thinking (see Section 2.5). We will now discuss some of the other models briefly.

Gemini 2.0 Pro

In February 2025, we released an experimental version of Gemini 2.0 Pro. At the time, it had the strongest coding performance of any model in the Gemini model family, as well as the best understanding and world knowledge. It also came with our largest context window at 2 million tokens, which enabled it to comprehensively analyze and understand vast amounts of information. For further information about Gemini 2.0 Pro, please see our earlier blog posts (Kavukcuoglu, 2025; Mallick and Kilpatrick, 2025).

Gemini 2.0 Flash Native Image Generation Model

In March 2025, we released an experimental version of Gemini 2.0 Flash Native Image Generation. It has brought to the users new capabilities as a result of a strong integration between the Gemini model and image-generation capabilities, enabling new experiences related to image generation & image editing via natural-language prompting. Capabilities such as multi-step conversational editing or interleaved text-image generation are very natural in such a setting, and horizontal transfer related to multi-language coverage immediately allowed such experiences to happen across all the languages supported by the Gemini models. Native image generation turns Gemini into a multimodal creation partner and enables Gemini to express ideas through both text and images, and to seamlessly move between the two. For further information about Gemini 2.0 Flash Native Image Generation, please see our earlier blog posts (Kampf and Brichtova, 2025; Sharon, 2025)

Gemini 2.5 Audio Generation

With Gemini 2.5, the Controllable TTS and Native Audio Dialog capabilities are available as separate options on AI Studio (Generate Media and Stream sections respectively). Our Gemini 2.5 Preview TTS Pro and Flash models support more than 80 languages with the speech style controlled by a free formatted prompt which can specify style, emotion, pace, etc, while also being capable of following finer-grained steering instructions specified in the transcript. Notably, Gemini 2.5 Preview TTS can generate speech with multiple speakers, which enables the creation of podcasts as used in NotebookLM Audio Overviews (Wang, 2024). Our Gemini 2.5 Flash Preview Native Audio Dialog model uses native audio generation, which enables the same level of style, pacing and accent control as available in our controllable TTS offering. Our dialog model supports tool use and function calling, and is available in more than 24 languages. With native audio understanding and generation capabilities, it can understand and respond appropriately to the user's tone. This model is also capable of understanding when to respond to the user, and when not to respond, ignoring background and non-device directed audio. Finally, we also offer an advanced 'Thinking' variant that effectively handles more complex queries and provides more robust and reasoned responses in exchange for some additional latency.

Gemini 2.5 Flash-Lite

In June 2025, we released an experimental version of Gemini 2.5 Flash-Lite (`gemini-2.5-flash-lite-preview-06-17`). It comes with the same capabilities that make Gemini 2.5 helpful, including the ability to turn thinking on at different budgets, connecting to tools like Google Search and code execution, support for multimodal inputs and a 1 million-token context length. Our goal was to provide an economical model class which provides ultra-low-latency capabilities and high throughput per dollar, echoing the initial release of 2.0 Flash-Lite (Google DeepMind, 2025b; Mallick and Kilpatrick, 2025).

Gemini 2.5 Pro Deep Think

To advance Gemini's capabilities towards solving hard reasoning problems, we developed a novel reasoning approach, called Deep Think, that naturally blends in parallel thinking techniques during response generation. Deep Think enables Gemini to creatively produce multiple hypotheses and carefully critique them before arriving at the final answer, achieving state-of-the-art performances in challenging benchmarks such as Olympiad math (USAMO 2025), competitive coding (LiveCodeBench), and multimodality (MMMU), see more details at (Doshi, 2025b). We announced Gemini 2.5 Deep Think at Google I/O and launched an experimental version to trusted testers and advanced users in June 2025.